

PAT-NO: JP403225445A
DOCUMENT-IDENTIFIER: JP 03225445 A
TITLE: LOAD DISPERSION STRIPING SYSTEM

PUBN-DATE: October 4, 1991

INVENTOR-INFORMATION:

| NAME | COUNTRY |
|----------------------|---------|
| NAKA, SEIICHIRO | |
| YASUGADAIRA, MICHICO | |

ASSIGNEE-INFORMATION:

| NAME | COUNTRY |
|---------------------------------|---------|
| NEC CORP | N/A |
| TOHOKU NIPPON DENKI SOFTWARE KK | N/A |

APPL-NO: JP02021341

APPL-DATE: January 30, 1990

INT-CL (IPC): G06F012/00 , G06F003/06

ABSTRACT:

PURPOSE: To prevent input and output from being congested on a specific storage device by determining storage devices to be assigned in predetermined order and updating a storage device control table according to the number of storage device and an assignment quantity which are calculated from the states of the respective storage devices stored in a storage device control table.

CONSTITUTION: The storage device control table 25 is referred to pieces of information on the input and output of the respective storage devices 4 to select a necessary number of storage devices in order from the device having the least input/output request quantity; when there are devices having the same quantity, the lengths of their queues are compared to select a shorter-length storage device and then a storage device having the largest data storage capacity.

Further, the storage device having the largest data storage capacity is selected in the storage device control table 25 and after the storage devices to be stored with data are determined, the data storage capacity of each storage device is updated. Actual input and output operation is performed by an input/output control part 3 and the length of the input/output queue and input/ output request quantity in the referred storage device control table 25 are updated. Consequently, a load is prevented from being concentrated on a specific storage device.

COPYRIGHT: (C)1991, JPO&Japio

⑫ 公開特許公報(A)

平3-225445

⑮ Int.Cl.⁵G 06 F 12/00
3/06

識別記号

3 0 1 B
3 0 1 J

庁内整理番号

8944-5B
7232-5B

⑬ 公開 平成3年(1991)10月4日

審査請求 未請求 請求項の数 1 (全6頁)

⑭ 発明の名称 負荷分散ストライピング方式

⑯ 特 願 平2-21341

⑰ 出 願 平2(1990)1月30日

⑱ 発 明 者 中 誠 一 郎 東京都港区芝5丁目33番1号 日本電気株式会社内
⑱ 発 明 者 安 ケ 平 達 子 宮城県仙台市青葉区中央4丁目6番1号 東北日本電気ソ
フトウェア株式会社内
⑲ 出 願 人 日 本 電 気 株 式 有 限 公 司 東京都港区芝5丁目7番1号
⑲ 出 願 人 東 北 日 本 電 気 ソ フ ト ウ ェ ア 株 式 有 限 公 司 宮城県仙台市青葉区中央4丁目6番1号
⑳ 代 理 人 弁 理 士 内 原 晋

明 細 書

発明の名称

負荷分散ストライピング方式

特許請求の範囲

入出力を制御する機構を備えたファイルシステムが、同一な入出力の処理速度を持つ複数の記憶装置上に、同一容量で分割して格納されている仮想パーティション方式における負荷分散ストライピング方式において、前記記憶装置の転送速度、処理効率を保つために必要なデータ量、要求される処理速度、記憶装置数を格納している仮想パーティション管理テーブルと、割り当て処理の際に、要求されているデータ割り当て量と前記仮想パーティション管理テーブルに格納された要求される処理速度から、前記要求される処理速度を満足するようにデータを分散させる記憶装置数と割り当て量を算出する手段と、各記憶装置の状態を示す格納できるデータ量、入出力に対する待ち行

列の長さ、入出力要求量とが格納されている記憶装置管理テーブルと、前記記憶装置管理テーブルに格納された前記各記憶装置の状態から前記算出された記憶装置数および割り当て量に従って予め定めた順に割り当てる記憶装置を決定するとともに前記記憶装置管理テーブルを更新する手段とを有することを特徴とする負荷分散ストライピング方式。

発明の詳細な説明

〔産業上の利用分野〕

本発明は、複数の記憶装置を資源とするコンピュータシステムのファイルシステムに関する。

〔従来の技術〕

従来のデータ格納方式は、データ量が多く一つの記憶装置に格納できない場合に、格納限度を超えた分を次の記憶装置に格納するような方式、または割り当てを行う各記憶装置の情報をシステムのテーブルに格納しておき、その情報を基に記憶装置を選択割り当てを行う方式、ユーザプログラ

ムでファイルを作成する際に使用する記憶装置を指定するような方式をとっていた。

(発明が解決しようとする課題)

上記従来のデータ格納方式では、ある特定の記憶装置にのみ入出力が集中したり、その反面で、全く入出力要求がない記憶装置が存在してしまうなど、特定の記憶装置に対する負荷が増加し、処理速度が低下するという問題点、また、ファイル作成時にストライピングの数が固定され、ユーザに物理的な記憶装置を意識させてしまうといった問題点がある。

(課題を解決するための手段)

本発明の負荷分散ストライピング方式は、各記憶装置の転送速度、処理効率を保つために必要なデータ量、要求される処理速度、記憶装置数を格納している仮想パーティション管理テーブルと、割り当て処理の際に、要求されているデータ割り当て量と前記仮想パーティション管理テーブルに格納された要求される処理速度から、前記要求される処理速度を満足するようにデータを分散させ

る記憶装置数と割り当て量を算出する手段と、各記憶装置の状態を示す格納できるデータ量、入出力に対する待ち行列の長さ、入出力要求量とが格納されている記憶装置管理テーブルと、前記記憶装置管理テーブルに格納された前記各記憶装置の状態から前記算出された記憶装置数および割り当て量に従って予め定められた順に割り当てる記憶装置を決定するとともに前記記憶装置管理テーブルを更新する手段とを有している。

(実施例)

次に本発明について図面を参照して説明する。

第1図は、本発明の構成を表す図である。

本発明は、ユーザプログラム1からの入出力に関する処理について、同一の処理速度と容量を持つn個からなる記憶装置4(DK1、DK2、…、DKn)と、それら記憶装置を制御する部分からなる。記憶装置4には既に仮想パーティション方式で生成されたファイルがパーティションVPとして格納されている。制御する部分はファイルシステム2としてファイル管理装置2と、

データ解放制御部22と、入出力制御部3と、データの割り当てを制御するデータ割当制御手段23と、仮想パーティション管理テーブル24と、記憶装置管理テーブル25で構成されている。

データ割当制御部23は、分散格納する記憶装置数と割り当て量を算出する手段231と、データを格納する記憶装置を決定する手段232とによって構成される。

第2図は、仮想パーティション管理テーブル24を示す概念図であり、パーティションVPの転送速度T、要求処理速度V、記憶装置の処理効率を保つために必要なデータ処理量U、記憶装置数Nが格納されている。

第3図は、記憶装置管理テーブル25を示す概念図であり、第2図のパーティションVPに対応する。各記憶装置DK1～DKnの格納可能なデータ量M1～Mn、入出力の待ち行列の長さQ1～Qn、入出力の待ち行列全体のデータ量L1～Lnが格納されている。ここで、入出力の待ち行

列の長さとして入出力の待ち行列全体のデータ量は入出力制御部3において、入出力要求発生時に更新されているものとする。

仮想パーティション方式によって構築されたファイルシステムVPに、ユーザプログラム1からXバイトの書き込み要求が発生する。ユーザプログラム1からの命令は、ファイル管理装置21でREAD制御部223、WRITE制御部221、DELETE制御部222に分けられ、それぞれの実行機構に制御が移る。

Xバイトの書き込み要求は、WRITE制御部221を経て、データ格納領域のデータ割当制御部23に制御が移される。データ割当制御部23における記憶装置数・割り当て量算出手段231では、書き込み要求に伴い発生するデータ割り当てが要求されると、パーティションVPに対応する仮想パーティション管理テーブル24を参照しながら、仮想パーティションに要求されている処理速度を満足するように分割するデータ量と記憶装置の数を算出する。

要求データXバイトを分割する際の手順は、次の通りである。

まず、仮想パーティション管理テーブル24よりパーティションVPの実際の記憶装置の転送速度がT、要求されている処理速度はV、記憶装置の処理効率を保つために必要なデータ処理量がUであることが分かる。

ここで、 $\alpha \times U$ バイト ($\alpha > 0$ 整数) を α 台の記憶装置にUバイトずつ同時にアクセスした場合、その転送速度V(α)は、

$$V(\alpha) = \alpha \times T \quad (式1)$$

で表すことができる。

よって、要求されたXバイトをq個の記憶装置に $p \times U$ バイトずつ平均的に分割格納するとすると、次の関係が成り立たなければならない。

$$X \leq p \times U \times q \quad (式2)$$

$$V \leq q \times T \quad (式3)$$

ただし、 $p \geq 1$ (整数)、 $1 \leq q \leq n$ (整数) とする。この関係式より、

$$q \geq V / T \quad (式4)$$

ただし、 $X < U$ の場合は、性能低下を避けるために常に1つの記憶装置にXバイト全てを割り当てることとする。

このようにして、Xバイトを q_1 個の記憶装置に

$$(p_1 \times U, p_1 \times U, \dots, X - (p_1 \times U \times (q_2 - 1)))$$

で分割する量が算出される。

次に、記憶装置選択手段232について説明する。

記憶装置数・割り当て量算出手段231より、 q_2 個の記憶装置に分割することを決定したが、

$$n \subset q_2 \quad (式12)$$

で算出することができる。

本発明においては、各記憶装置に対する負荷をできるだけ均等化するために、記憶装置管理テーブル25より、各々の記憶装置の入出力に対する情報を参照し、入出力要求量の最も少ないものから順に q_2 個の記憶装置を選択することとする。同一のものがあつた場合は、待ち行列の長さを比

$$p \geq X / (q \times U) \quad (式5)$$

が成り立つ。

式4を満足する最小の記憶装置数qを q_1 で表すと

$$q_1 = [V / T] \quad (式6)$$

([] はガウス記号である。)

この q_1 を基に式5は次のように変形される。

$$p \geq X / (q_1 \times U) \quad (式7)$$

式7を満足する最小のデータ割り当て量pを p_1 で表すと、

$$p_1 = [X / (q_1 \times U)] \quad (式8)$$

また、

$$p_1 \geq X / (q_1 \times U) \quad (式9)$$

ゆえに、

$$q_1 \geq X / (p_1 \times U) \quad (式10)$$

よって、実際に割り当てる記憶装置数を q_2 とすると、

$$q_2 = [X / (p_1 \times U)] \leq q_1 \quad (式11)$$

で求めることができる。

較し、その短いもの、次に格納可能なデータ容量の最も大きい記憶装置から選択することとする。記憶装置管理テーブル25より格納可能なデータ容量の最も大きい記憶装置から選択することとする。データを格納する記憶装置が決定した後で、各記憶装置のデータ格納可能量の更新を行う。実際の入出力は入出力制御部3で行われ、前記参照した記憶装置管理テーブル25の入出力待ち行列の長さが入出力要求量が更新される。

次にXバイトのデータ解放要求がユーザプログラム1から発生した場合について述べる。

この場合は、ファイル管理装置21のDELETE制御部222を経て、データ解放制御部22に制御が移る。ここで、各記憶装置に対応する記憶装置管理テーブル25のデータ格納可能量の更新を行い、入出力制御部3で実際のデータ領域の解放処理が行われる。

最後に、Xバイトのデータ読み込み要求がユーザプログラム1から発生した場合について述べる。

ファイル管理装置21のREAD制御部233を経て、入出力制御部3において、実際のデータ読み込みが行われるが、ここで、データが割り付けられている各記憶装置に対応する記憶装置管理テーブル25の入出力待ち行列の長さ、入出力要求量が更新される。

次に、ここまで述べた負荷分散パーティション方式について、具体的な数字をあげてその処理を説明する。

第4図、第5図は仮想パーティションVPに対する仮想パーティション管理テーブル24と記憶装置管理テーブル25を示す概念図である。

仮想パーティションVPは、4台の記憶装置(DK1, DK2, DK3, DK4)で構成されており、その処理速度は100で皆同一である。この性能を満足するために必要なデータ量が20であり、ユーザから要求されている処理速度は150である。この環境の中に、ユーザより140のデータ割り当て要求が発生したとする。

まず、記憶装置数・割り当て量案出手段231

により、割り当てるべき装置台数 q_1 及びその割り当て量 p_1 を計算する。式6、式8より、

$$q_1 = [150 / 100] = 2$$

$$p_1 = [140 / (2 \times 20)] = 4$$

よって、式12より実際に割り当てる台数 q_2 は、

$$q_2 = [140 / (3 \times 20)] = 2$$

以上より、要求量140は2台の記憶装置に(80, 60)と分散させることとする。

次に、第5図より各記憶装置の入出力負荷状態を参照して、2台の記憶装置を決定する。

組合せとしては装置が4台あるので、 $4C_2 = 6$ 通り考えられるが、まず、入出力要求量(必要格納量20を単位として表す)の最も少ない順で考えると、

$$DK1 < DK2 = DK3 < DK4$$

しかし、DK2とDK3の入出力要求量が同一であるので、次に、入出力待ち行列の長さを参照すると、

$$DK1 < DK3 < DK2 < DK4$$

よって、DK1とDK3を2台の記憶装置として選択することになる。

したがって、この例の場合、要求量140は次のようにストライピングは、

$$(DK1, DK2, DK3, DK4) \\ = (60, 0, 60, 0)$$

この分散を施したあとの記憶装置テーブルを第6図に示す。

前記データ割り当てを実施した後で、ユーザより100のデータ解放要求が発生したとすると、ファイル管理装置21において、実際データが割り付けられている記憶装置を得る。ここでは、 $(DK2, DK4) = (60, 40)$ の割合で割り付けられていたと仮定する。

データ解放要求はDELETE制御部222を経て、データ解放制御部22に制御が移り、第6図で表される記憶装置管理テーブル25の格納可能量を更新する。

$$DK2 \cdots 480 \rightarrow 480 + 60 = 540$$

$$DK4 \cdots 500 \rightarrow 500 + 40 = 540$$

この解放を施したあとの記憶装置管理テーブル25を第7図に示す。

(発明の効果)

以上説明したように本発明は、ユーザに物理的な記憶装置を意識させずに、ファイル書き込み時に動的にストライピングの数を決定し、実際の入出力の負荷状態を基に、ユーザが求める入出力の性能を実現し、特定の記憶装置に入出力が集中することを防ぐよう効率的な記憶装置の活用とデータの割り当てを行う効果がある。

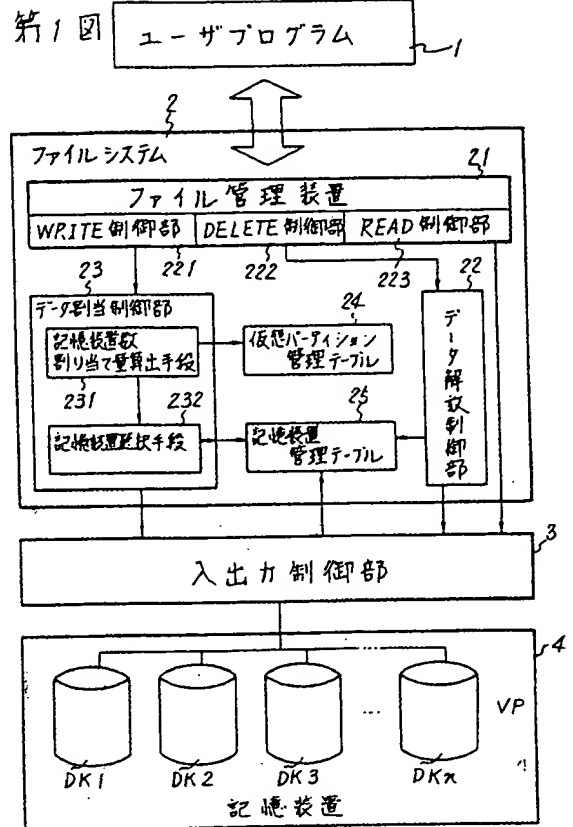
図面の簡単な説明

第1図は本発明の一実施例を示す図、第2図は第1図に示す仮想パーティション管理テーブルを示す図、第3図は第1図に示す記憶装置管理テーブルを示す図、第4図は具体的な数字を使って仮想パーティション管理テーブルを示した図、第5図は具体的な数字を使って記憶装置管理テーブルを示した図、第6図は分散を実施したあと更新した記憶装置管理テーブルを示した図、第7図は解

放を実施したあと更新した記憶装置管理テーブルを示した図である。

1…ユーザプログラム、2…ファイルシステム、3…入出力制御部、4…記憶装置、21…ファイル管理装置、22…データ解放制御部、23…データ割当制御部、24…仮想パーティション管理テーブル、25…記憶装置管理テーブル、231…記憶装置数・割り当て量算出手段、232…記憶装置選択手段。

代理人 弁理士 内 原 晋



第2図

24 仮想パーティション管理テーブル

| パーティション | VP |
|---------|----|
| 転送速度 | T |
| 必要格納量 | U |
| 要求処理速度 | V |
| 装置数 | N |

第4図

24 仮想パーティション管理テーブル

| パーティション | VP |
|---------|-----|
| 転送速度 | 100 |
| 必要格納量 | 20 |
| 要求処理速度 | 150 |
| 装置数 | 4 |

第3図

25 記憶装置管理テーブル

| 記憶装置ID | DK1 | DK2 | ... | DKn |
|----------|-----|-----|-----|-----|
| 格納可能量 | M1 | M2 | ... | Mn |
| 入出力待ち行列長 | Q1 | Q2 | ... | Qn |
| 入出力要求量 | L1 | L2 | ... | Ln |

第5図

25 記憶装置管理テーブル

| 記憶装置ID | DK1 | DK2 | DK3 | DK4 |
|----------|-----|-----|-----|-----|
| 格納可能量 | 600 | 480 | 550 | 500 |
| 入出力待ち行列長 | 1 | 3 | 2 | 4 |
| 入出力要求量 | 2 | 4 | 4 | 6 |

第 6 図

25 記憶装置管理テーブル

| 記憶装置ID | DK1 | DK2 | DK3 | DK4 |
|---------|-----|-----|-----|-----|
| 格納可能量 | 520 | 480 | 490 | 500 |
| 入出力待行列長 | 2 | 3 | 3 | 4 |
| 入出力要求量 | 6 | 4 | 7 | 6 |

第 7 図

25 記憶装置管理テーブル

| 記憶装置ID | DK1 | DK2 | DK3 | DK4 |
|---------|-----|-----|-----|-----|
| 格納可能量 | 520 | 540 | 490 | 540 |
| 入出力待行列長 | 2 | 3 | 3 | 4 |
| 入出力要求量 | 6 | 4 | 7 | 6 |